Lecture 5. Online Convex Optimization

Advanced Optimization (Fall 2023)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Online Learning

- Online Convex Optimization

- Convex Functions

- Strongly Convex Functions

- Exp-concave Functions

A Brief Review of Statistical Learning

The fundamental goal of (supervised) learning: ***Risk Minimization (RM)***,

$$\min_{h \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}}[\ell(h(\mathbf{x}), y)],$$

where

- h denotes the hypothesis (model) from the hypothesis space $\mathcal{H}$.

- $(\mathbf{x}, y)$ is an instance chosen from a unknown distribution $\mathcal{D}$.

- $\ell(h(\mathbf{x}), y)$ denotes the loss of using hypothesis h on the instance $(\mathbf{x}, y)$.

A Brief Review of Statistical Learning

Given a loss function f and distribution $\mathcal{D}$, the *expected risk* of predictor h is

$$R(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h(\mathbf{x}), y)].$$

In practice, we can only access to a sample set $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$. Thus, the following *empirical risk* is naturally defined:

$$\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h(\mathbf{x}_i), y_i).$$

A Brief Review of Statistical Learning

- A successful story : characterization of sample complexity

  ☐ excess risk bound

  $$R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h) \leq \mathcal{O}\left(\frac{1}{\sqrt{m}}\right).$$

  ☐ generalization error bound

  $$\hat{R}_S(\hat{h}) - R(\hat{h}) \leq \mathcal{O}\left(\frac{1}{\sqrt{m}}\right).$$

Offline Towards Online Learning

- Traditional statistical machine learning

  - The training data are available *offline*

  - Learning model is trained based on the offline data in a *batch* setting

- Online learning scenario

  - In real applications, data are in the form of *stream*

  - New data are being collected all the time: after observing a new data point, the model should be *online updated* at a constant cost

A Formulation of Online Learning

- We introduce a ***game-theoretic view*** to model online learning.

- Online learning is formulated as a repeated game between
    - ***Player***: essentially the learner, or you can think as the "learning model"
    - ***Environments***: an abstraction of all factors evaluating the model.

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{w}_t \in \mathcal{W}$;

(2) and simultaneously environments pick an online function $f_t : \mathcal{W} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{w}_t)$, observes some information about f_t and updates the model.

Online Learning: Formulation

At each round $t = 1, 2, \cdots$

   (1) the player first picks a model $\mathbf{w}_t \in \mathcal{W}$;

   (2) and simultaneously environments pick an online function $f_t : \mathcal{W} \to \mathbb{R}$;

   (3) the player suffers loss $f_t(\mathbf{w}_t)$, observes some information about f_t and updates the model.

- An example of online function $f_t : \mathcal{W} \mapsto \mathbb{R}$.

Considering the task of **online classification**, we have

   (i) the loss $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}$, and

   (ii) the hypothesis function $h : \mathcal{W} \times \mathcal{X} \mapsto \hat{\mathcal{Y}}$.

$\Longrightarrow$

$$f_t(\mathbf{w}) = \ell(h(\mathbf{w}; \mathbf{x}_t), y_t)$$
$$= \ell(\mathbf{w}^\top \mathbf{x}_t, y_t) \quad \textit{for simplicity}$$

Online Learning: Formulation

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{w}_t \in \mathcal{W}$;

(2) and simultaneously environments pick an online function $f_t : \mathcal{W} \rightarrow \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{w}_t)$, observes some information about f_t and updates the model.

Spam filtering
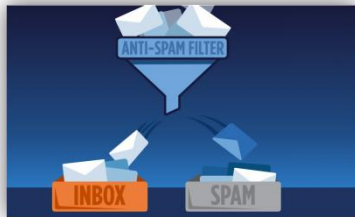
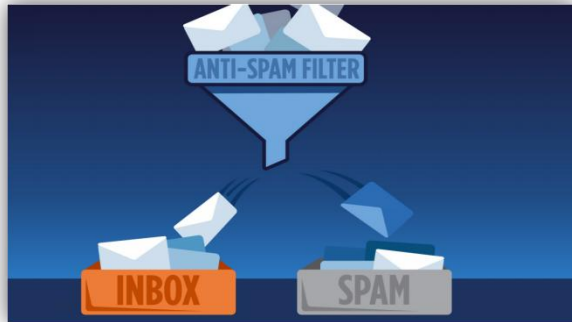(1) Player submits a spam classifier $\mathbf{w}_t$

(2) A mail is revealed whether it is a spam

(3) Player suffers loss $f_t(\mathbf{w}_t)$ and updates model

Applications

spam detection (online classification/regression): At each time $t = 1, 2, \ldots$

- receive an email $\mathbf{x}_t \in \mathbb{R}^d$;

- predict whether it is a spam $\hat{y}_t \in \{-1, +1\}$;

- see its true label $y_t \in \{-1, +1\}$.

aggregating weather prediction (the expert problem): At each day $t = 1, 2, \ldots$

- obtain temperature predictions from N models;

- make the final prediction by randomly following a model according to the probability $p_t \in \Delta_N$;

- on the next day observe the loss of each model $f_t \in [0, 1]^N$.

Performance Measure

- Recall in the statistical learning: *risk*

$$R(h) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\ell(h(\mathbf{x}),y)\big].$$

- In online learning: *sequential risk*

$$\hat{R}\big(\{\mathbf{w}_t\}_{t=1}^{T}\big) \triangleq \sum_{t=1}^{T} f_t(\mathbf{w}_t) = \sum_{t=1}^{T} \ell(\mathbf{w}_t^{\top}\mathbf{x}_t, y_t).$$

meaning: cumulative loss of models trained on growing data stream $S_t = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)\}$.

Performance Measure

- In offline learning, we use ***excess risk*** as measure for $\hat{h}$:

$$R(\hat{h}) - \min_{h \in \mathcal{H}} R(h)$$

$$= \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}}[\ell(\mathbf{w}^\top \mathbf{x}, y)] - \min_{\mathbf{w} \in \mathcal{W}} \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}}[\ell(\mathbf{w}^\top \mathbf{x}, y)]$$

simply using a linear model **w** *to parametrize the hypothesis h*

- In online learning, we define ***regret*** as measure for sequence $\{\mathbf{w}_t\}_{t=1}^T$:

$$R(\{\mathbf{w}_t\}_{t=1}^T) - \min_{\mathbf{w} \in \mathcal{W}} R(\{\mathbf{w}\}_{t=1}^T)$$

$$= \sum_{t=1}^T \ell(\mathbf{w}_t^\top \mathbf{x}_t, y_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^T \ell(\mathbf{w}^\top \mathbf{x}_t, y_t)$$

Performance Measure

- In offline learning, we use ***excess risk*** as measure for $\hat{h}$:

$$R(\hat{h}) - \min_{h \in \mathcal{H}} R(h)$$

$$= \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}}[\ell(\mathbf{w}^\top \mathbf{x}, y)] - \min_{\mathbf{w} \in \mathcal{W}} \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}}[\ell(\mathbf{w}^\top \mathbf{x}, y)]$$

*simply using a linear model **w** to parametrize the hypothesis h*

- In online learning, we define ***regret*** as measure for sequence $\{\mathbf{w}_t\}_{t=1}^T$:

$$R(\{\mathbf{w}_t\}_{t=1}^T) - \min_{\mathbf{w} \in \mathcal{W}} R(\{\mathbf{w}\}_{t=1}^T)$$

$$= \sum_{t=1}^T \ell(\mathbf{w}_t^\top \mathbf{x}_t, y_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^T \ell(\mathbf{w}^\top \mathbf{x}_t, y_t)$$

$$\text{Regret}_T = \sum_{t=1}^T f_t(\mathbf{w}_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^T f_t(\mathbf{w})$$

benchmark performance with the offline model (optimal in hindsight)

Another View of Regret

- Ultimate goal: minimize the ***cumulative loss*** $\sum_{t=1}^{T} f_t(\mathbf{w}_t)$

- The cumulative loss highly depends on the loss function itself, so we need a benchmark:

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w})$$

- We hope the regret be sub-linear dependence with T.

$$\frac{\text{Regret}_T}{T} \to 0 \text{ as } T \to \infty$$ ***Hannan Consistency***

Compared with Statistical Offline Learning

• Memory efficient

• Do not need i.i.d. assumption
  • the environment can be even adversarial
  • typically, the regret analysis does not need concentration

• Strictly harder than statistical learning
  • under non-i.i.d. assumption
  • online to batch conversion

Online Learning: Is it (provably) solvable?

- In general, the online learning formulation is *hard* to solve.

At each round $t = 1, 2, \cdots$

   (1) the player first picks a model $\mathbf{w}_t \in \mathcal{W}$;

   (2) and simultaneously environments pick an online function $f_t : \mathcal{W} \to \mathbb{R}$;

   (3) the player suffers loss $f_t(\mathbf{w}_t)$, observes some information about f_t and updates the model.
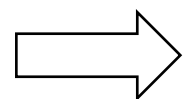
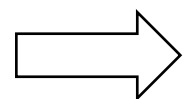$\implies$ A Trackable Case: ***Online Convex Optimization***

requiring feasible domain and online functions to be *convex*

Online Learning: Is it (provably) solvable?

- In general, the online learning formulation is *hard* to solve.

At each round $t = 1, 2, \cdots$

  (1) the player first picks a model $\boxed{\mathbf{w}_t \in \mathcal{W}}$;

  (2) and simultaneously environments pick an online function $\boxed{f_t : \mathcal{W} \to \mathbb{R}}$;

  (3) the player suffers loss $f_t(\mathbf{w}_t)$, observes some information about f_t and updates the model.

$\Longrightarrow$ A Trackable Case: ***Online Convex Optimization***

requiring feasible domain and online functions to be *convex*

Online Convex Optimization

- OCO framework
  - feasible domain is a convex set
  - online functions are convex

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a convex set $\mathcal{X} \subseteq \mathbb{R}^d$;

(2) and environments pick an online convex function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

From this point forward, we use $\mathbf{x}$ (and $\mathcal{X}$) instead of $\mathbf{w}$ (and $\mathcal{W}$) for consistency with opt. language.

OCO: Different Setups
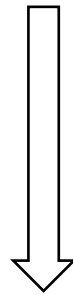
At each round $t = 1, 2, \cdots$

   (1) the player first picks a model $\mathbf{x}_t$ from a convex set $\mathcal{X} \subseteq \mathbb{R}^d$;

   (2) and environments pick an online convex function $f_t : \mathcal{X} \to \mathbb{R}$;

   (3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.
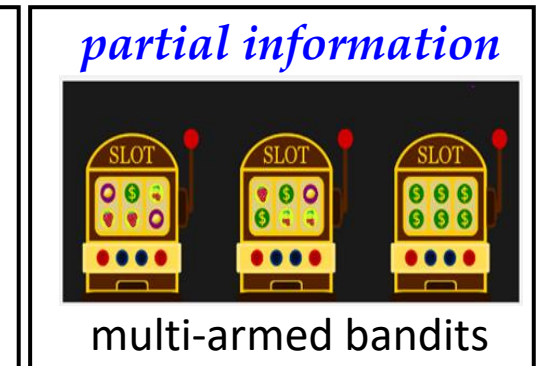
on the feedback information:

- *full information*: observe entire f_t (or at least gradient $\nabla f_t(\mathbf{x}_t)$)

- *partial information (bandits)*: observe function value $f_t(\mathbf{x}_t)$ only

less information

full information

horse racing

partial information

multi-armed bandits

OCO: Different Setups
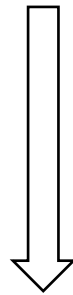
At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a convex set $\mathcal{X} \subseteq \mathbb{R}^d$;

(2) and environments pick an online convex function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

on the difficulty of environments:

- stochastic setting

- adversarial setting $\begin{cases} \text{oblivious} \\ \text{adaptive} \\ \text{(non-oblivious)} \end{cases}$

less restricted but harder
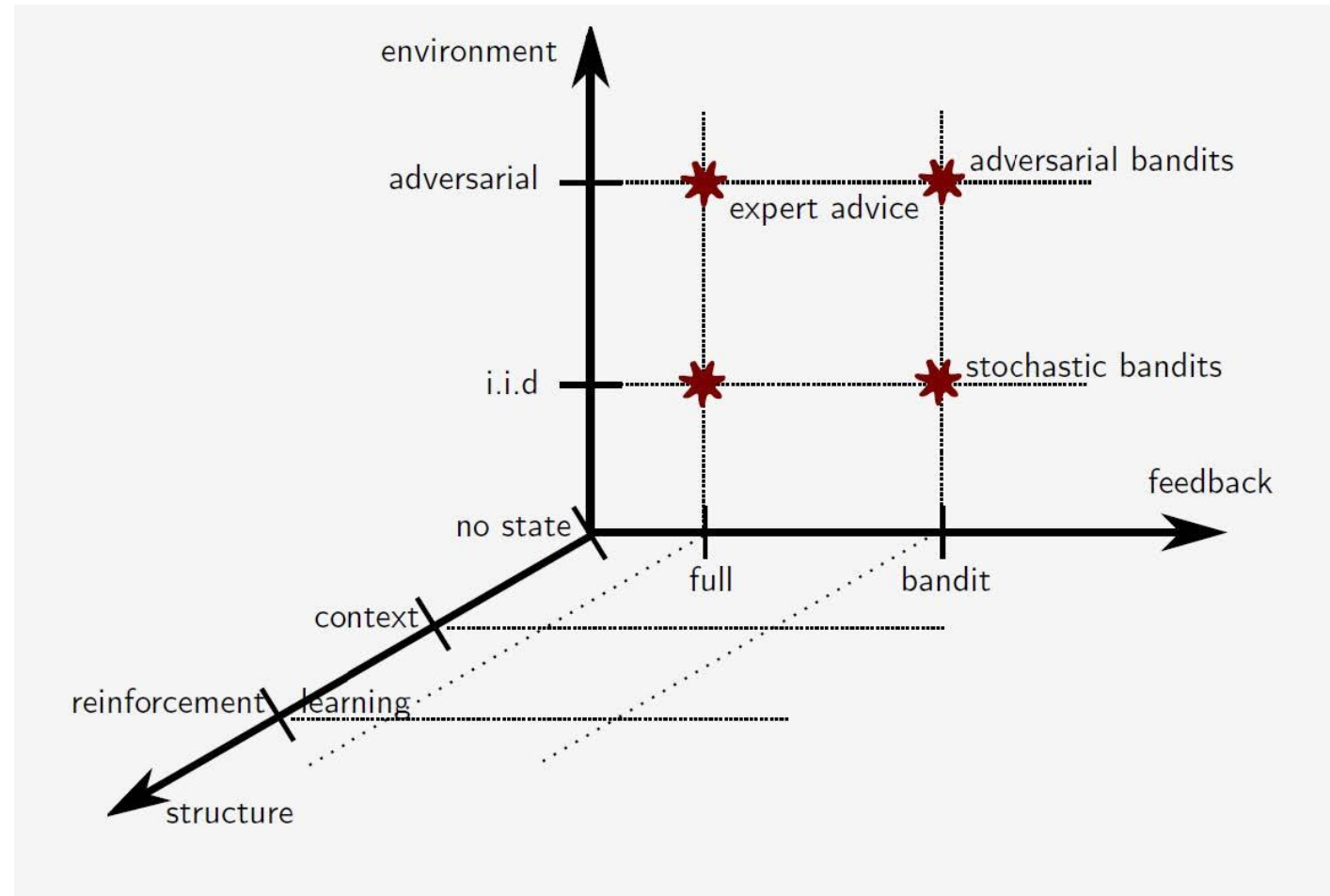
oblivious adversary

examination

adaptive adversary

interview

The Space of Online Learning Problems

Yevgeny Seldin. The Space of Online Learning Problems, ECML-PKDD, Porto, Portugal, 2015.

Online Learning

- Full-information setting:
  - Online Convex Optimization
  - Prediction with Expert Advice
  - ...

- Partial-information setting:
  - Multi-Armed Bandits
  - Linear Bandits/Parametric Bandits
  - Bandit Convex Optimization
  - ...

Online Learning

- **Full-information setting:**
  - Online Convex Optimization
  - Prediction with Expert Advice
  - ...

- Partial-information setting:
  - Multi-Armed Bandits
  - Linear Bandits/Parametric Bandits
  - Bandit Convex Optimization
  - ...

History: Two-Player Zero-Sum Games

Theory of repeated games

James Hannan
(1922–2010)

David Blackwell
(1919–2010)

Learning to play a game (1956)

Play a game repeatedly against a possibly suboptimal opponent

Zero-sum 2-person games played more than once

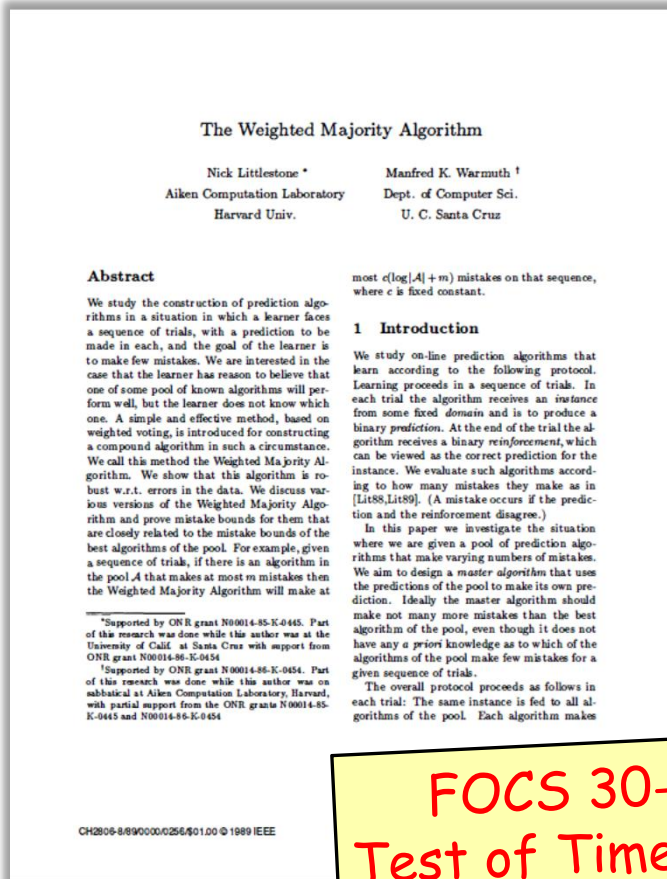|	1	2	...	M
1	$\ell(1,1)$	$\ell(1,2)$	...	
2	$\ell(2,1)$	$\ell(2,2)$	...	
⋮	⋮	⋮	⋱	
N				

$N \times M$ known loss matrix
- Row player (player) has N actions
- Column player (opponent) has M actions

For each game round $t = 1, 2, \ldots$
- Player chooses action i_t and opponent chooses action y_t
- The player suffers loss $\ell(i_t, y_t)$ ($=$ gain of opponent)

Player can learn from opponent's history of past choices $y_1, \ldots, y_{t-1}$

Nicolo Cesa-Bianchi, Online Learning and Online Convex Optimization. Tutorial at the Simons Institute. 2017.

History: Prediction with Expert Advice

Manfred Warmuth

UC Santa Cruz

FOCS 30-year
Test of Time Award!

Nick Littlestone and Manfred K. Warmuth.
"The Weighted Majority Algorithm." FOCS 1989: 256-261.
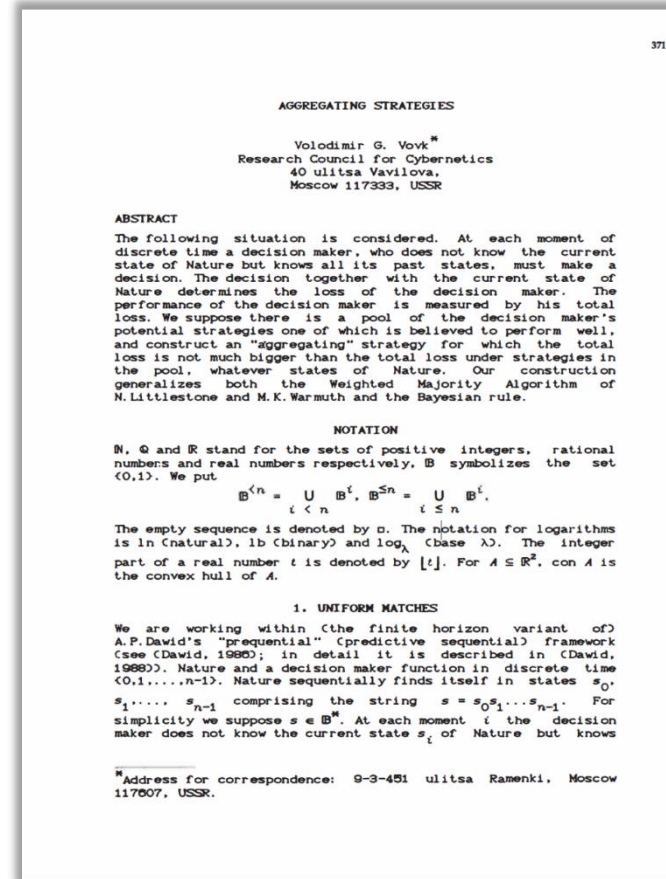
Volodimir G. Vovk
Royal Holloway,
University of London

Volodimir G. Vovk. "Aggregating
Strategies." COLT 1990: 371-383.

Online Convex Optimization

- Convex Functions

- Strongly Convex Functions

- Exponentially Concave Functions

Online Convex Optimization

- Convex Functions

- Strongly Convex Functions

- Exponentially Concave Functions

Online Optimization with Convex Functions

Definition 2 (Convex Function). A function $f : \mathcal{X} \mapsto \mathbb{R}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$\forall \alpha \in [0, 1], f((1 - \alpha)\mathbf{x} + \alpha \mathbf{y}) \leq (1 - \alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

Equivalently, if f is differentiable, we have that $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^{\top}(\mathbf{y} - \mathbf{x}).$$

Online Gradient Descent

Online Gradient Descent (OGD)

At each round $t = 1, 2, \cdots$

1. the player first picks a model $\mathbf{x}_t \in \mathcal{X}$;

2. and simultaneously environments pick a ***convex loss function*** $f_t : \mathcal{X} \to \mathbb{R}$;

3. the player suffers loss $f_t(\mathbf{x}_t)$, observes the information (loss) f_t and update the model according to $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right]$.

- $\Pi_{\mathcal{X}}[\mathbf{y}] = \arg\min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|_2$ denotes the Euclidean projection onto the feasible set $\mathcal{X}$.

- This belongs to the full-information setting, so player can access the gradient $\nabla f_t(\mathbf{x}_t)$.

   But actually the gradient is the only required, so it's also called *gradient-feedback* OCO model.

OGD: Regret Analysis

- The following assumptions are required for standard analysis.

Assumption 1 (Convexity). The feasible set $\mathcal{X}$ is closed and convex in Euclidean space, and $f_1, \ldots, f_T$ are convex functions.

Assumption 2 (Bounded Decision Set). The diameter of the set $\mathcal{X}$ is upper bounded by D, i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \|\mathbf{x} - \mathbf{y}\| \leq D$.

Assumption 3 (Bounded Gradient). The norm of the subgradients of f is upper bounded by G, i.e., $\|\nabla f(\mathbf{x})\| \leq G$ for all $\mathbf{x} \in \mathcal{X}$.

OGD: Regret Analysis

Theorem 3 (Regret bound for OGD). *Under Assumptions 1, 2 and 3, online gradient descent (OGD) with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$ for $t \in [T]$ guarantees:*

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2}GD\sqrt{T} = \mathcal{O}(\sqrt{T}).$$

The First Gradient Descent Lemma

Lemma 1. *Suppose that f is proper, closed and convex; the feasible domain $\mathcal{X}$ is nonempty, closed and convex. Let $\{\mathbf{x}_t\}_{t=1}^{T}$ be the sequence generated by the gradient descent method. Then for any $\mathbf{u} \in \mathcal{X}^{\star}$ and $t \geq 0$,*

$$\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 \leq \|\mathbf{x}_t - \mathbf{u}\|^2 - 2\eta_t(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) + \eta_t^2\|\nabla f_t(\mathbf{x}_t)\|^2.$$

Proof:

$$\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 = \|\Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t\nabla f_t(\mathbf{x}_t)] - \mathbf{u}\|^2 \quad \text{(GD)}$$

$$\leq \|\mathbf{x}_t - \eta_t\nabla f_t(\mathbf{x}_t) - \mathbf{u}\|^2 \quad \text{(Pythagoras Theorem)}$$

$$= \|\mathbf{x}_t - \mathbf{u}\|^2 - 2\eta_t\langle\nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u}\rangle + \eta_t^2\|\nabla f_t(\mathbf{x}_t)\|^2$$

$$\leq \|\mathbf{x}_t - \mathbf{u}\|^2 - 2\eta_t(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) + \eta_t^2\|\nabla f_t(\mathbf{x}_t)\|^2$$

$$\text{(convexity: } f_t(\mathbf{x}_t) - f_t(\mathbf{u}) = f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \langle\nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u}\rangle) \quad \square$$

Proof for OGD Regret Bound

Proof: We use the first gradient descent lemma to analyze online gradient descent.

Lemma 1. *Suppose that f is proper, closed and convex; the feasible domain $\mathcal{X}$ is nonempty, closed and convex. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by the gradient descent method. Then for any $\mathbf{u} \in \mathcal{X}^\star$ and $t \geq 0$,*

$$\textcolor{red}{\|\mathbf{x}_{t+1} - \mathbf{u}\|^2} \leq \|\mathbf{x}_t - \mathbf{u}\|^2 - 2\eta_t(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) + \eta_t^2 \|\nabla f_t(\mathbf{x}_t)\|^2.$$

By Lemma 1,

$$2(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \leq \frac{\|\mathbf{x}_t - \mathbf{u}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{u}\|^2}{\eta_t} + \eta_t G^2$$

Proof for OGD Regret Bound

Proof: By setting $\eta_t = \frac{D}{G\sqrt{t}}$ (with $\frac{1}{\eta_0} := 0$), summing over T:

$$2\left(\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{u})\right) \leq \sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \mathbf{u}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{u}\|^2}{\eta_t} + G^2 \sum_{t=1}^{T} \eta_t \qquad \text{(GD lemma)}$$

$$\leq \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{u}\|^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t \qquad (\|\mathbf{x}_{T+1} - \mathbf{u}\|^2 \geq 0)$$

$$\leq D^2 \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=1}^{T} \eta_t$$

$$\leq D^2 \frac{1}{\eta_T} + G^2 \sum_{t=1}^{T} \eta_t \qquad (\eta_t = \frac{D}{G\sqrt{t}} \text{ and } \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T})$$

$$\leq 3DG\sqrt{T}. \qquad \qquad \square$$
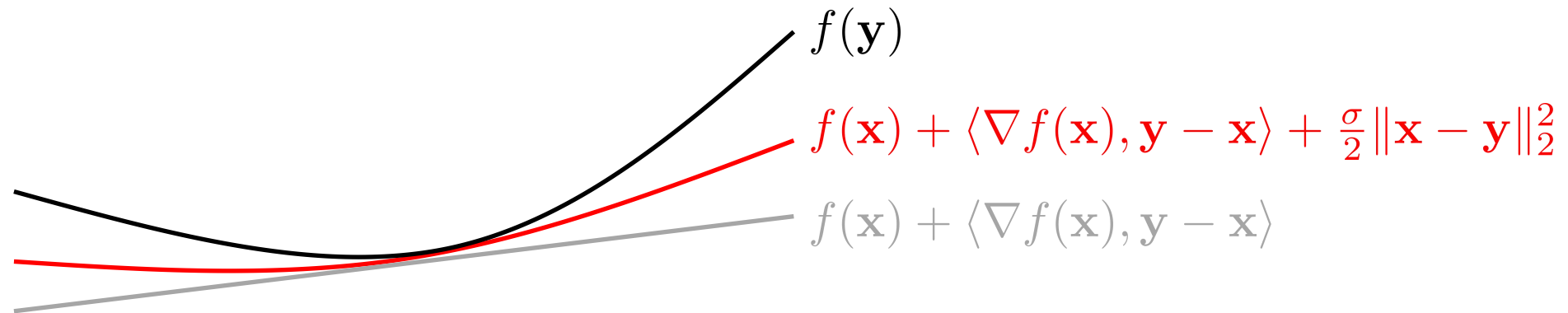
Online Convex Optimization

• Convex Functions

• **Strongly Convex Functions**

• Exponentially Concave Functions

OCO with Strongly Convex Functions

Definition 3 (Strong Convexity). A function f is σ-strongly convex if, for any $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\sigma}{2} \|\mathbf{y} - \mathbf{x}\|^2,$$

or equivalently, $\nabla^2 f(\mathbf{x}) \succeq \alpha I$.

$f(\mathbf{y})$

$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$

$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$

OGD for Strongly Convex Loss

Online Gradient Descent (OGD)

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t \in \mathcal{X}$;

(2) and simultaneously environments pick a ***strongly convex loss*** $f_t : \mathcal{X} \rightarrow \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes the information (loss) f_t and update the model according to $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t)\right]$.

- The learning rate for strongly convex OGD is set as $\eta_t = \frac{1}{\sigma t}$.

OGD for Strongly Convex Loss

Theorem 4 (Regret bound for strongly-convex functions). *Under Assumption 1 and Assumption 3, for σ-strongly convex loss functions, online gradient descent with step sizes $\eta_t = \frac{1}{\sigma t}$ achieves the following guarantee*

$$\text{Regret}_T \leq \frac{G^2}{2\sigma}(1 + \log T) = \mathcal{O}(\log T).$$

• Strongly convex case compared with convex case: $\mathcal{O}(\log T)$ vs. $\mathcal{O}(\sqrt{T})$

• A caveat is that we now don't need Assumption 2 (bounded domain).

OCO with Strongly Convex Functions

Proof: we start by extending *the first GD lemma* to strongly convex case.

Strongly convex case:

$$\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 \leq \|\mathbf{x}_t - \mathbf{u}\|^2 - 2\eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle + \eta_t^2 \|\nabla f_t(\mathbf{x}_t)\|^2$$

$$\leq \|\mathbf{x}_t - \mathbf{u}\|^2 - 2\eta_t \left(f_t(\mathbf{x}_t) - f_t(\mathbf{u}) {\color{red}+ \frac{\sigma}{2}\|\mathbf{x}_t - \mathbf{u}\|^2} \right) + \eta_t^2 \|\nabla f_t(\mathbf{x}_t)\|^2$$

(strong convexity: $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) + \frac{\sigma}{2}\|\mathbf{x}_t - \mathbf{u}\|^2 \leq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle$)

$$\leq {\color{red}(1 - \sigma\eta_t)} \|\mathbf{x}_t - \mathbf{u}\|^2 - 2\eta_t \left(f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \right) + \eta_t^2 \|\nabla f_t(\mathbf{x}_t)\|^2$$

$$\implies f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \frac{\eta_t^{-1} - \sigma}{2} \|\mathbf{x}_t - \mathbf{u}\|^2 - \frac{\eta_t^{-1}}{2} \|\mathbf{x}_{t+1} - \mathbf{u}\|^2 + \frac{\eta_t G^2}{2} \quad \text{(rearranging)}$$

OCO with Strongly Convex Functions

Proof: $\quad f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \dfrac{\eta_t^{-1} - \sigma}{2}\|\mathbf{x}_t - \mathbf{u}\|^2 - \dfrac{\eta_t^{-1}}{2}\|\mathbf{x}_{t+1} - \mathbf{u}\|^2 + \dfrac{\eta_t G^2}{2}$

Summing from $t = 1$ to T, setting $\eta_t = \frac{1}{\sigma t}$ $\left(\text{define } \frac{1}{\eta_0} := 0\right)$:

$$2\sum_{t=1}^{T}\big(f_t(\mathbf{x}_t) - f_t(\mathbf{u})\big) \leq \sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{u}\|^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma\right) + G^2\sum_{t=1}^{T}\eta_t \quad \left(\tfrac{1}{\eta_0} := 0\right)$$

$$= 0 + G^2\sum_{t=1}^{T}\frac{1}{\sigma t} \quad \left(\tfrac{1}{\eta_0} \triangleq 0,\ \|\mathbf{x}_{T+1} - \mathbf{u}\|^2 \geq 0,\ \tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} - \sigma = 0\right)$$

$$\leq \frac{G^2}{\sigma}(1 + \log T). \qquad\qquad \square$$

Online Convex Optimization

- Convex Functions

- Strongly Convex Functions

- **Exponentially Concave Functions**

Comparison of (Strongly) Convex Problems

Convex Problem
Property: $f_t(\mathbf{x}) \geq f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$
OGD: $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \dfrac{1}{\sqrt{t}} \nabla f_t(\mathbf{x}_t) \right]$
$\text{Regret}_T \leq \dfrac{3}{2} G D \sqrt{T}$

Strongly Convex Problem
Property: $f_t(\mathbf{y}) \geq f_t(\mathbf{x}) + \nabla f_t(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$ $+ \dfrac{\sigma}{2} \|\mathbf{y} - \mathbf{x}\|^2$
OGD: $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \dfrac{1}{\sigma t} \nabla f_t(\mathbf{x}_t) \right]$
$\text{Regret}_T \leq \dfrac{G^2}{2\sigma} (1 + \log T)$

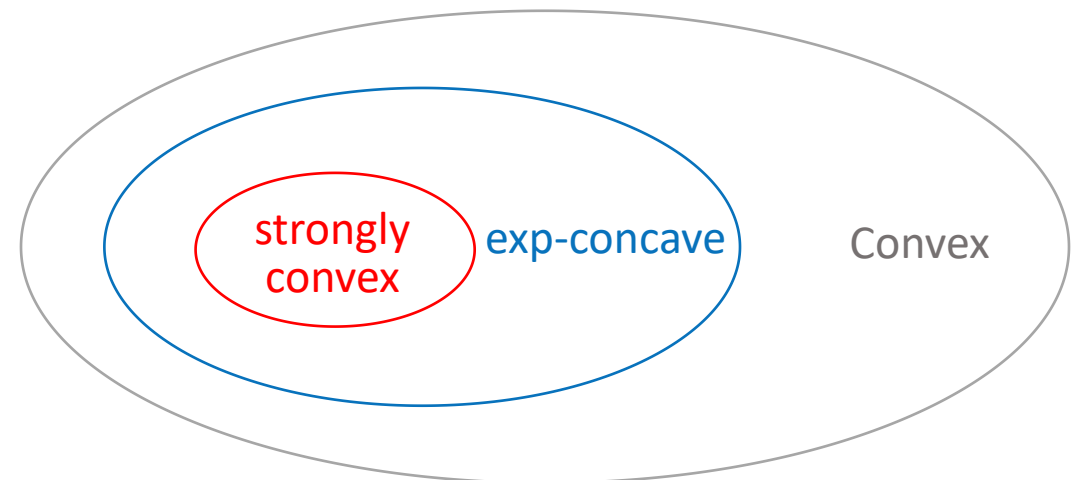Can we explore more function class with a regret rate faster than $\sqrt{T}$?

Exponentially-concave Function

Definition 2 (Exp-concavity). A convex function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is defined to be α-exp-concave over $\mathcal{X} \subseteq \mathbb{R}^d$ if the function g is concave, where $g : \mathcal{X} \mapsto \mathbb{R}$ is defined as

$$g(\mathbf{x}) = e^{-\alpha f(\mathbf{x})}.$$

Directly employ OGD: Regret bound $\mathcal{O}(\sqrt{T})$.

But actually we can get a **tighter** bound!

strongly convex — exp-concave — Convex

An Example for Exp-concave Learning

- Universal Portfolio Selection

  - a total of d stocks in the stock market.

  - each round, the player chooses stocks by a distribution $\mathbf{x}_t \in \Delta_d$.

  - the market returns the price ratio $\boldsymbol{\theta}_t$ between iter t and $t+1$,

$$\boldsymbol{\theta}_t(i) = \frac{\text{price of stock}_i \text{ at time } t+1}{\text{price of stock}_i \text{ at time } t}$$

  which means that our final wealth W_T will be: $W_T = W_1 \cdot \prod_{t=1}^{T} \boldsymbol{\theta}_t^\top \mathbf{x}_t$

$\Longrightarrow$ Our goal is to maximize our wealth at time T.

An Example for Exp-concave Learning

- Universal Portfolio Selection

  - we hope to maximize the logarithm of W_T

  $$\log \frac{W_T}{W_1} = \sum_{t=1}^{T} \log \boldsymbol{\theta}_t^\top \mathbf{x}_t$$

  - using OCO framework,

  $$\boxed{f_t(\mathbf{x}) = \log(\boldsymbol{\theta}_t^\top \mathbf{x})}$$

  At each round $t = 1, 2, \cdots$

  (1) the player first picks a model $\mathbf{x}_t \in \Delta_d$;

  (2) and simultaneously environments pick an online function $f_t : \mathcal{X} \to \mathbb{R}$;

  (3) the player get a **gain** $f_t(\mathbf{x}_t) = \log(\boldsymbol{\theta}_t^\top \mathbf{x}_t)$, observes f_t and updates the model.

  - Goal: $\quad \text{Regret}_T = \max_{\mathbf{x}^\star \in \Delta_d} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) - \sum_{t=1}^{T} f_t(\mathbf{x}_t)$

  $\boxed{\textbf{online function is exp-concave}}$

Exponential-concave Function

Lemma 3 (Property of Exp-concavity). *Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be an α-exp-concave function, and D, G denote the diameter of $\mathcal{X}$ and a bound on the (sub)gradients of f respectively. The following holds for all $\gamma \leq \frac{1}{2} \min \left\{ \frac{1}{GD}, \alpha \right\}$ and all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^{\top}(\mathbf{x} - \mathbf{y}) {\color{red} + \frac{\gamma}{2}(\mathbf{x} - \mathbf{y})^{\top} \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^{\top}(\mathbf{x} - \mathbf{y})}.$$

Proof. Recall that f is α-exp-concave if and only if $e^{-\alpha f(\mathbf{x})}$ is concave.

As $2\gamma \leq \alpha$, $e^{-2\gamma f(\mathbf{x})} = \left(e^{-\alpha f(\mathbf{x})} \right)^{2\gamma/\alpha}$ is also concave and thus is 2γ-exp-concave.

$$e^{-2\gamma f(\mathbf{x})} - e^{-2\gamma f(\mathbf{y})} \leq \left\langle \mathbf{x} - \mathbf{y}, -2\gamma e^{-2\gamma f(\mathbf{y})} \nabla f(\mathbf{y}) \right\rangle.$$

<div align="right">(concavity)</div>

Exponential-concave Function

Lemma 3 (Property of Exp-concavity). *Let $f : \mathcal{X} \to \mathbb{R}$ be an α-exp-concave function, and D, G denote the diameter of $\mathcal{X}$ and a bound on the (sub)gradients of f respectively. The following holds for all $\gamma \leq \frac{1}{2} \min \left\{ \frac{1}{GD}, \alpha \right\}$ and all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:*

$$f(\mathbf{x}) \;\geq\; f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) {\color{red}+ \frac{\gamma}{2} (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})}.$$

Proof. Dividing $e^{-2\gamma f(\mathbf{y})}$ at both sides achieves

$$f(\mathbf{y}) - f(\mathbf{x}) \leq \frac{1}{2\gamma} \log \left(1 + \boxed{2\gamma \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{y}) \rangle} \right).$$

Our constructive condition $\gamma \leq \frac{1}{2} \min \left\{ \frac{1}{GD}, \alpha \right\}$ ensures $|2\gamma \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{y}) \rangle| \leq 1$,

$$f(\mathbf{y}) - f(\mathbf{x}) \leq \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{y}) \rangle - \frac{\gamma}{2} \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{y}) \rangle^2$$

$$(\log(1 + x) \leq x - \tfrac{1}{4}x^2) \text{ holds for } (|x| \leq 1) \qquad \square$$

A Comparison of Different Curvatures

- Convex
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$$

- Strongly Convex
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \textcolor{red}{+ \frac{\sigma}{2} \|\mathbf{x} - \mathbf{y}\|_2^2}$$

- Exponentially Concave
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \textcolor{red}{+ \frac{\gamma}{2} (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})}$$
$$= f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \textcolor{red}{+ \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|_{\nabla f(\mathbf{y}) \nabla f(\mathbf{y})^\top}^2}$$

Exponential-concave Function

Lemma 3 (Property of Exp-concavity). *Let $f : \mathcal{X} \to \mathbb{R}$ be an α-exp-concave function, and D, G denote the diameter of $\mathcal{X}$ and a bound on the (sub)gradients of f respectively. The following holds for all $\gamma \leq \frac{1}{2} \min \left\{ \frac{1}{GD}, \alpha \right\}$ and all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:*

$$f(\mathbf{x}) \; \geq \; f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{\gamma}{2} (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$$

$$= f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|^2_{\nabla f(\mathbf{y}) \nabla f(\mathbf{y})^\top}$$

Algorithmic intuition:

- For convex loss, we use 2-norm to encode the structure of the space.

- Can we exploit *local structures* of exp-concave loss to improve the regret?

Intuition

- Convex $$f_t(\mathbf{x}) \geq f_t(\mathbf{x}_t) + \nabla f_t(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t)$$

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right] \qquad \text{OGD with } \eta_t = \mathcal{O}(1/\sqrt{t})$$

- Strongly convex $$f_t(\mathbf{x}) \geq f_t(\mathbf{x}_t) + \nabla f_t(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t) + \frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2$$

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right] \qquad \text{OGD with } \eta_t = \mathcal{O}(1/t)$$

- Exp-concave $$f_t(\mathbf{x}) \geq f_t(\mathbf{x}_t) + \nabla f_t(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_t\|_{\nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^\top}^2$$

$\Longrightarrow$ We may still GD update, but the step size should be "***data-dependent***".

Intuitively, step size should be stretched *hetrogeneously* *in different directions,*
being smaller when $\nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^\top$ *is "larger".*

ONS for Exp-concave Function

Online Newton Step

Input: parameters $\gamma, \varepsilon > 0$, matrix $A_0 = \varepsilon I_d$

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$;

(2) and simultaneously environments pick an ***exp-concave loss function*** $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes the information (loss) f_t and update:

Update $A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^\top$

Update $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\| \mathbf{x} - \left(\mathbf{x}_t - \frac{1}{\gamma} A_t^{-1} \nabla f_t(\mathbf{x}_t) \right) \right\|_{A_t}^2$

ONS: In a View of Proximal Gradient

Convex Problem	**Exp-concave Problem**
Property: $f_t(\mathbf{x}) \geq f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$	Property: $f_t(\mathbf{x}) \geq f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$ $+ \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|^2_{\nabla f_t(\mathbf{y}) \nabla f_t(\mathbf{y})^\top}$
OGD: $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \frac{1}{\sqrt{t}} \nabla f_t(\mathbf{x}_t) \right]$	ONS: $A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^\top$ $\mathbf{x}_{t+1} = \Pi^{A_t}_{\mathcal{X}} \left[\mathbf{x}_t - \frac{1}{\gamma} A_t^{-1} \nabla f_t(\mathbf{x}_t) \right]$
Proximal type update: $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{x} - \mathbf{x}_t\|_2^2$	Proximal type update: $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_t\|_{A_t}^2$

ONS: In a View of Proximal Gradient

Proof.

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}^{A_t} \left[\mathbf{x}_t - \frac{A_t^{-1}}{\gamma} \mathbf{g}_t \right] \quad (\mathbf{g}_t \triangleq \nabla f_t(\mathbf{x}_t))$$

$$= \arg\min_{\mathbf{x} \in \mathcal{X}} \left(\mathbf{x} - \mathbf{x}_t + \frac{A_t^{-1}}{\gamma} \mathbf{g}_t \right)^{\top} A_t \left(\mathbf{x} - \mathbf{x}_t + \frac{A_t^{-1}}{\gamma} \mathbf{g}_t \right)$$

$$= \arg\min_{\mathbf{x} \in \mathcal{X}} \left(\mathbf{x} - \mathbf{x}_t + \frac{A_t^{-1}}{\gamma} \mathbf{g}_t \right)^{\top} \left(A_t \mathbf{x} - A_t \mathbf{x}_t + \frac{\mathbf{g}_t}{\gamma} \right)$$

$$= \arg\min_{\mathbf{x} \in \mathcal{X}} \left(\mathbf{x} - \mathbf{x}_t \right)^{\top} A_t \left(\mathbf{x} - \mathbf{x}_t \right) + \cancel{\left(A^{-1} \right)^{\top} \mathbf{g}_t^{\top} \mathbf{g}_t}$$

$$\qquad\qquad\qquad + 2 \frac{\mathbf{g}_t^{\top} \left(\mathbf{x} - \mathbf{x}_t \right)}{\gamma} \; \text{(constant)}$$

$$= \arg\min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \mathbf{g}_t \rangle + \frac{\gamma}{2} \| \mathbf{x} - \mathbf{x}_t \|_{A_t}^2$$

Exp-concave Problem

Property: $f_t(\mathbf{x}) \;{\color{red}\geq}\; f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^{\top} (\mathbf{x} - \mathbf{y})$

$${\color{red} + \frac{\gamma}{2} \| \mathbf{x} - \mathbf{y} \|_{\nabla f_t(\mathbf{y}) \nabla f_t(\mathbf{y})^{\top}}^2}$$

ONS: $\quad A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^{\top}$

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}^{A_t} \left[\mathbf{x}_t - \frac{1}{\color{red}\gamma} A_t^{-1} \nabla f_t(\mathbf{x}_t) \right]$$

Proximal type update:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{\color{red}\gamma}{\color{red}2} \| \mathbf{x} - \mathbf{x}_t \|_{A_t}^2$$

ONS for Exp-concave Function

Theorem 5. *Under Assumptions 1, 2 and 3, for α-exp-concave online functions, the ONS algorithm with parameters $\gamma = \frac{1}{2}\min\left\{\frac{1}{GD}, \alpha\right\}$ and $\varepsilon = \frac{1}{\gamma^2 D^2}$ (recall that the initial matrix is $A_0 = \varepsilon I_d$) guarantees*

$$\mathrm{Regret}_T \le \mathcal{O}\left(\left(\frac{1}{\alpha} + GD\right)d\log T\right) = \mathcal{O}(d\log T),$$

where d is the dimension of the feasible domain $\mathcal{X} \subseteq \mathbb{R}^d$.

Proof

$\bullet\ A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top$

$\bullet\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\| \mathbf{x} - (\mathbf{x}_t - \frac{1}{\gamma}A_t^{-1}\mathbf{g}_t) \right\|_{A_t}^2$

Extending *the first GD lemma* to ***exp-concave case***:

Proof. We use norm induced by A_t instead of 2-norm.

$$\|\mathbf{x}_{t+1} - \mathbf{u}\|_{A_t}^2 = \left\| \Pi_{\mathcal{X}}^{A_t}\left[\mathbf{x}_t - \frac{1}{\gamma}A_t^{-1}\nabla f_t(\mathbf{x}_t)\right] - \mathbf{u} \right\|_{A_t}^2 \qquad (\Pi_{\mathcal{X}}^{A}[\mathbf{y}] \triangleq \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|_A^2)$$

$$\leq \left\| \mathbf{x}_t - \frac{1}{\gamma}A_t^{-1}\nabla f_t(\mathbf{x}_t) - \mathbf{u} \right\|_{A_t}^2 \qquad\qquad \begin{array}{c}(A_t \text{ is semidefinite matrix})\\ (\text{Pythagoras theorem})\end{array}$$

$$= \left(\mathbf{x}_t - \frac{1}{\gamma}A_t^{-1}\nabla f_t(\mathbf{x}_t) - \mathbf{u}\right)^\top A_t \left(\mathbf{x}_t - \frac{1}{\gamma}A_t^{-1}\nabla f_t(\mathbf{x}_t) - \mathbf{u}\right)$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{definition of } \|\cdot\|_{A_t}^2)$$

$$= \left(\mathbf{x}_t - \mathbf{u} - \frac{1}{\gamma}A_t^{-1}\nabla f_t(\mathbf{x}_t)\right)^\top \left(A_t(\mathbf{x}_t - \mathbf{u}) - \frac{1}{\gamma}\nabla f_t(\mathbf{x}_t)\right)$$

Proof

Extending *the first GD lemma* to ***exp-concave case***:

Proof.

$$\|\mathbf{x}_{t+1} - \mathbf{u}\|_{A_t}^2 = \left(\mathbf{x}_t - \mathbf{u} - \frac{1}{\gamma} A_t^{-1} \nabla f_t(\mathbf{x}_t) \right)^\top \left(A_t(\mathbf{x}_t - \mathbf{u}) - \frac{1}{\gamma} \nabla f_t(\mathbf{x}_t) \right)$$

$$= (\mathbf{x}_t - \mathbf{u})^\top A_t (\mathbf{x}_t - \mathbf{u}) - \frac{2}{\gamma} \nabla f_t(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{u}) + \frac{1}{\gamma^2} \nabla f_t(\mathbf{x}_t)^\top A_t^{-1} \nabla f_t(\mathbf{x}_t)$$

$$\leq \|\mathbf{x}_t - \mathbf{u}\|_{A_t}^2 - \frac{2}{\gamma} \left(f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \right) + \frac{1}{\gamma^2} \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$

$$- (\mathbf{x}_t - \mathbf{u})^\top \nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{u})$$

(Exp-concave: $f_t(\mathbf{x}) \geq f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{\gamma}{2} (\mathbf{x} - \mathbf{y})^\top \nabla f_t(\mathbf{y}) \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$)

Proof

Proof. $\|\mathbf{x}_{t+1} - \mathbf{u}\|_{A_t}^2$

$$\leq \|\mathbf{x}_t - \mathbf{u}\|_{A_t}^2 - \frac{2}{\gamma}\big(f_t(\mathbf{x}_t) - f_t(\mathbf{u})\big) - \|\mathbf{x}_t - \mathbf{u}\|_{\nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top}^2 + \frac{1}{\gamma^2}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$

$$\Longrightarrow f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \frac{\gamma}{2}\|\mathbf{x}_t - \mathbf{u}\|_{A_t}^2 - \frac{\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{u}\|_{A_t}^2 - \frac{\gamma}{2}\|\mathbf{x}_t - \mathbf{u}\|_{\nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top}^2 + \frac{1}{2\gamma}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$

(rearranging)

Summing from $t = 1$ to T, by telescoping:

$$\sum_{t=1}^{T}\big(f_t(\mathbf{x}_t) - f_t(\mathbf{u})\big) \leq \frac{\gamma}{2}\sum_{t=1}^{T}\Big(\|\mathbf{x}_t - \mathbf{u}\|_{A_t}^2 - \|\mathbf{x}_t - \mathbf{u}\|_{A_{t-1}}^2\Big) + \frac{1}{2\gamma}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$

$$+ \frac{\gamma}{2}\|\mathbf{x}_1 - \mathbf{u}\|_{A_0}^2 - \frac{\gamma}{2}\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{u}\|_{\nabla f(\mathbf{x}_t)\nabla f(\mathbf{x}_t)^\top}^2$$

cancellation

$$\leq \frac{\gamma}{2}\|\mathbf{x}_1 - \mathbf{u}\|_{A_0}^2 + \frac{1}{2\gamma}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$

$(A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top)$

Proof

Proof.
$$\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t) - f_t(\mathbf{u})\right) \leq \frac{\gamma}{2}\left\|\mathbf{x}_1 - \mathbf{u}\right\|_{A_0}^2 + \frac{1}{2\gamma}\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t)\right\|_{A_t^{-1}}^2$$

By the definition that $A_0 \triangleq \varepsilon I_d$, $\varepsilon = \frac{1}{\gamma^2 D^2}$ and the diameter $\left\|\mathbf{x}_1 - \mathbf{u}\right\|_2^2 \leq D^2$:

$$\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t) - f_t(\mathbf{u})\right) \leq \frac{\gamma}{2}\left(\mathbf{x}_1 - \mathbf{u}\right)^{\top} A_0 \left(\mathbf{x}_1 - \mathbf{u}\right) + \frac{1}{2\gamma}\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t)\right\|_{A_t^{-1}}^2$$

$$\leq \frac{1}{2\gamma} + \frac{1}{2\gamma}\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t)\right\|_{A_t^{-1}}^2 .$$

Next, we bound the term $\sum_{t=1}^{T}\left\|\nabla f_t(\mathbf{x}_t)\right\|_{A_t^{-1}}^2$.

Proof

Proof. Next, we bound the term $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$.

Lemma 4 (Elliptical Potential Lemma). *For any sequence $\{X_1, \ldots, X_T\} \in \mathbb{R}^{d \times T}$, suppose $U_0 = \lambda I, U_t = U_{t-1} + X_t X_t^{\top}$, and $\|X_t\|_2 \leq L$, then*

$$\sum_{t=1}^{T} \|X_t\|_{U_t^{-1}}^2 \leq d \log \left(1 + \frac{L^2 T}{\lambda d} \right)$$

Proof. $\quad U_{t-1} = U_t - X_t X_t^{\top} = U_t^{\frac{1}{2}} \left(I - U_t^{-\frac{1}{2}} X_t X_t^{\top} U_t^{-\frac{1}{2}} \right) U_t^{\frac{1}{2}}$ (definition of U_t)

$$\det(U_{t-1}) = \det(U_t) \det \left(I - U_t^{-\frac{1}{2}} X_t X_t^{\top} U_t^{-\frac{1}{2}} \right) \quad \text{(determinant on both side)}$$

Proof

Proof. Next, we bound the term $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$.

Lemma 5. *For any $\mathbf{v} \in \mathbb{R}^d$, we have*

$$\det\left(I - \mathbf{v}\mathbf{v}^{\top}\right) = 1 - \|\mathbf{v}\|_2^2$$

Proof.

   (i) $\left(I - \mathbf{v}\mathbf{v}^{\top}\right)\mathbf{v} = \left(1 - \|\mathbf{v}\|_2^2\right)\mathbf{v}$, therefore, $\mathbf{v}$ is its eigenvector with $\left(1 - \|\mathbf{v}\|_2^2\right)$ as eigenvalue;

   (ii) $\left(I - \mathbf{v}\mathbf{v}^{\top}\right)\mathbf{v}^{\perp} = \mathbf{v}^{\perp}$, therefore, $\mathbf{v}^{\perp} \perp \mathbf{v}$ is its eigenvector with 1 as the eigenvalue. $\qquad \square$

Proof

Proof. Next, we bound the term $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$.

Lemma 4 (Elliptical Potential Lemma). *For any sequence $\{X_1, \ldots, X_T\} \in \mathbb{R}^{d \times T}$, suppose $U_0 = \lambda I, U_t = U_{t-1} + X_t X_t^\top$, and $\|X_t\|_2 \leq L$, then*

$$\sum_{t=1}^{T} \|X_t\|_{U_t^{-1}}^2 \leq d \log \left(1 + \frac{L^2 T}{\lambda d} \right)$$

Proof. $\det(U_{t-1}) = \det(U_t) \det \left(I - U_t^{-\frac{1}{2}} X_t X_t^\top U_t^{-\frac{1}{2}} \right) = \det(U_t) \left(1 - \left\| U_t^{-\frac{1}{2}} X_t \right\|_2^2 \right)$

(by Lemma 5)

$$\implies \|X_t\|_{U_t^{-1}}^2 = \left\| U_t^{-\frac{1}{2}} X_t \right\|_2^2 = 1 - \frac{\det(U_{t-1})}{\det(U_t)} \qquad \text{(rearranging, } U \text{ is a symmetric matrix)}$$

Proof

Proof. Next, we bound the term $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$.

Lemma 4 (Elliptical Potential Lemma). *For any sequence $\{X_1, \ldots, X_T\} \in \mathbb{R}^{d \times T}$, suppose $U_0 = \lambda I, U_t = U_{t-1} + X_t X_t^\top$, and $\|X_t\|_2 \leq L$, then*

$$\sum_{t=1}^{T} \|X_t\|_{U_t^{-1}}^2 \leq d \log \left(1 + \frac{L^2 T}{\lambda d} \right)$$

Proof.

$\implies \quad \sum_{t=1}^{T} X_t^\top U_t^{-1} X_t = \sum_{t=1}^{T} \left(1 - \frac{\det(U_{t-1})}{\det(U_t)} \right) \leq \sum_{t=1}^{T} \log \frac{\det(U_t)}{\det(U_{t-1})} \quad (\forall x > 0, 1 - x \leq -\log x)$

$\quad = \log \frac{\det(U_T)}{\det(U_0)} = d \log \left(1 + \frac{L^2 T}{\lambda d} \right) \quad \square \quad \begin{array}{l} \text{Tr}(U_T) \leq \text{Tr}(U_0) + L^2 T = \lambda d + L^2 T \\ \implies \det(U_T) \leq (\lambda + L^2 T/d)^d \end{array}$

Proof

Proof. Next, we bound the term $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$.

> **Lemma 4** (Elliptical Potential Lemma). *For any sequence* $\{X_1, \ldots, X_T\} \in \mathbb{R}^{d \times T}$, *suppose* $U_0 = \lambda I, U_t = U_{t-1} + X_t X_t^{\top}$, *and* $\|X_t\|_2 \leq L$, *then*
>
> $$\sum_{t=1}^{T} \|X_t\|_{U_t^{-1}}^2 \leq d \log \left(1 + \frac{L^2 T}{\lambda d} \right)$$

Therefore, by Lemma 4, we have

$$\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2 \leq d \log \left(1 + \frac{D^2 T}{\varepsilon d} \right).$$

Proof

Proof. To conclude,

$$\sum_{t=1}^{T} \left(f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \right) \le \underbrace{\frac{\gamma}{2} \left\| \mathbf{x}_1 - \mathbf{u} \right\|_{A_0}^2}_{\substack{\le \frac{1}{2\gamma} \\ \text{(bounded domain)}}} + \underbrace{\frac{1}{2\gamma} \sum_{t=1}^{T} \left\| \nabla f_t(\mathbf{x}_t) \right\|_{A_t^{-1}}^2}_{\substack{\le \frac{d}{2\gamma} \log \left(1 + \frac{D^2 T}{\varepsilon d} \right) \\ \text{(elliptical potential lemma)}}}.$$

Recall that $\gamma = \frac{1}{2} \min \left\{ \frac{1}{GD}, \alpha \right\}$ and $\varepsilon = \frac{1}{\gamma^2 D^2}$,

$$\text{Regret}_T \le \mathcal{O} \left(\left(\frac{1}{\alpha} + GD \right) d \log T \right). \qquad \square$$

Lower Bounds

- A natural question: whether previous regret can be improved?

- Lower bound argument:

  minimax bound: smallest possible worst-case regret of any algorithm:

$$\min_{\mathcal{A}} \max_{\ell_1,\ldots,\ell_T} \text{Regret}_T$$

Theorem 7 (Lower Bound for OCO). *Any algorithm for online convex optimization incurs $\Omega(DG\sqrt{T})$ regret in the worst case. This is true even if the cost functions are generated from a fixed stationary distribution.*

Lower Bounds

> **Theorem 7** (Lower Bound for OCO). *Any algorithm for online convex optimization incurs $\Omega(DG\sqrt{T})$ regret in the worst case. This is true even if the cost functions are generated from a fixed stationary distribution.*

Proof Sketch.

Construct a **'hard'** environment:

- Binary classification, loss functions in each iteration are chosen at random

- Similar results can be obtained for strongly convex and exp-concave cases

Comparison

	Algorithm	**Upper Bound**	**Lower Bound**
Convex	OGD	$\mathcal{O}(\sqrt{T})$	$\Omega(\sqrt{T})$
σ-Strongly Convex	OGD	$\mathcal{O}(\frac{\log T}{\sigma})$	$\Omega(\frac{\log T}{\sigma})$
α-Exp-concave	ONS	$\mathcal{O}(\frac{d \log T}{\alpha})$	$\Omega(\frac{d \log T}{\alpha})$

Back to Exp-concave Learning

- Universal Portfolio Selection

Algorithm	Regret	Runtime (per round)
Universal Portfolios	$d\log(T)$	$d^4 T^{14}$
Online Gradient Descent	$G_2\sqrt{T}$	d
Exponentiated Gradient	$G_\infty\sqrt{T\log(d)}$	d
Online Newton Step (ONS)	$G_\infty d\log(T)$	d^2+generalized projection on Δ_d
Soft-Bayes	$\sqrt{dT\log(d)}$	d
Ada-BARRONS	$d^2\log^4(T)$	$d^{2.5}T$
BISONS	$d^2\log^2(T)$	$\mathrm{poly}(d)$
AdaMix+DONS	$d^2\log^5(T)$	d^3
VB-FTRL	$d\log(T)$	$d^2 T$

[COLT 2020 Open Problem]

⇨ still an important open problem: efficiency and optimality

Connection with Offline Learning

- Application to Stochastic Optimization

- Online-to-Batch Conversion

Application to Stochastic Optimization

- Consider the following *convex optimization* problem:

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

- Stochastic optimization method

Computational oracle: only access *noisy* gradient oracle, namely, $\mathbf{g}(\mathbf{x})$, such that

$$\mathbb{E}[\mathbf{g}(\mathbf{x})] = \nabla f(\mathbf{x}), \text{ and } \mathbb{E}[\|\mathbf{g}(\mathbf{x})\|^2] \leq G^2$$

for some $G > 0$.

Example (large-scale opt.). Given dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, ERM optimizes

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{m} \ell\big(h(\mathbf{x}_i), y_i\big) \implies$$

full gradient computation requires a pass of *all data*

stochastic method only uses a *mini batch* at each round

Stochastic Gradient Descent

- Consider the following *convex optimization* problem:

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

Algorithm 2 Stochastic Gradient Descent

Input: noisy gradient oracle $\mathbf{g}(\cdot)$, step sizes $\{\eta_t\}$

1: **for** $t = 1, \ldots, T$ **do**
2:    Obtain noisy gradient $\mathbf{g}(\mathbf{x}_t)$
3:    Update the model $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \eta_t \mathbf{g}(\mathbf{x}_t) \right]$
4: **end for**
5: **return** $\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t$

$$\mathbb{E}[\mathbf{g}(\mathbf{x})] = \nabla f(\mathbf{x})$$

$$\mathbb{E}[\|\mathbf{g}(\mathbf{x})\|^2] \leq G^2$$

Stochastic Gradient Descent

> **Theorem 7** (Convergence of SGD). *Suppose the domain $\mathcal{X} \subseteq \mathbb{R}^d$ has a diameter $D > 0$, and the noisy gradient oracle is unbiased and variance bounded by G^2. SGD with step size $\eta_t = \frac{D}{G\sqrt{t}}$ guarantees*
>
> $$\mathbb{E}[f(\bar{\mathbf{x}}_T)] - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \leq \frac{3GD}{2\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right),$$
>
> *where $\bar{\mathbf{x}}_T \triangleq \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t$ is the output of the SGD algorithm.*

Proof of SGD Convergence

Proof. First, we rephrase SGD from lens of ***online convex optimization***.

To see this, we define linear function $h_t(\mathbf{x}) \triangleq \mathbf{g}_t^\top \mathbf{x}$, where $\mathbf{g}_t = \mathbf{g}(\mathbf{x}_t)$.

Claim: deploying OGD over the online functions $\{h_t(\mathbf{x})\}$ is equivalent to SGD proposed in the earlier page.

$$\textbf{OGD}: \mathbf{x}_{t+1} = \Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \nabla h_t(\mathbf{x}_t)]$$
$$= \Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \mathbf{g}(\mathbf{x}_t)]$$

Algorithm 2 Stochastic Gradient Descent

Input: noisy gradient oracle $\mathbf{g}(\cdot)$, step sizes $\{\eta_t\}$
1: **for** $t = 1, \ldots, T$ **do**
2:     Obtain noisy gradient $\mathbf{g}(\mathbf{x}_t)$
3:     Update the model $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \mathbf{g}(\mathbf{x}_t)]$
4: **end for**
5: **return** $\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t$

Proof of SGD Convergence

Proof.

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_T)\right] - f(\mathbf{x}^\star) \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_t)\right] - f(\mathbf{x}^\star)$$

$(\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}))$

(Jensen's inequality)

$$\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star)\right]$$

(convexity)

$$= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)\right]$$

Theorem 3 (Regret bound for OGD). *Under Assumption 1, 2 and 3, online gradient descent (OGD) with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$ for $t \in [T]$ guarantees:*

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}}\sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2}GD\sqrt{T}.$$

Proof:
   Define $\boldsymbol{\xi}_t \triangleq \nabla f(\mathbf{x}_t) - \mathbf{g}_t$. We know $\mathbb{E}[\boldsymbol{\xi}_t \mid \mathbf{x}_t] = 0$.
   We have $\mathbb{E}[\nabla f(\mathbf{x}_t)^\top \mathbf{x}_t] = \mathbb{E}[\boldsymbol{\xi}_t^\top \mathbf{x}_t] + \mathbb{E}[\mathbf{g}_t^\top \mathbf{x}_t]$
   $\qquad \mathbb{E}[\boldsymbol{\xi}_t^\top \mathbf{x}_t] = \mathbb{E}[\mathbb{E}[\boldsymbol{\xi}_t^\top \mathbf{x}_t \mid \mathbf{x}_t]] = \mathbb{E}[\mathbb{E}[\mathbb{E}[\boldsymbol{\xi}_t \mid \mathbf{x}_t]^\top \mathbf{x}_t \mid \mathbf{x}_t]] = 0.$
   Therefore, we have proved that $\mathbb{E}[\nabla f(\mathbf{x}_t)^\top \mathbf{x}_t] = \mathbb{E}[\mathbf{g}_t^\top \mathbf{x}_t].$
   Similar argument shows $\mathbb{E}[\nabla f(\mathbf{x}_t)^\top \mathbf{x}] = \mathbb{E}[\mathbf{g}_t^\top \mathbf{x}]$ for any *fixed* $\mathbf{x}$. $\square$

Proof of SGD Convergence

Proof.

$$\mathbb{E}\left[f(\overline{\mathbf{x}}_T)\right] - f(\mathbf{x}^\star) \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_t)\right] - f(\mathbf{x}^\star)$$

$(\mathbf{x}^\star = \arg\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}))$

(Jensen's inequality)

$$\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star)\right]$$

(convexity)

$$= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)\right]$$

$$= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} h_t(\mathbf{x}_t) - h_t(\mathbf{x}^\star)\right]$$

(definition of $f_t(\cdot)$)

$$\leq \frac{\text{Regret}_T}{T}$$

(SGD = OGD over $\{f_t(\cdot)\}$)

(regret bound of OGD)

$$\leq \frac{3GD}{2\sqrt{T}}$$

$\square$

(regret of OGD algorithm)

Theorem 3 (Regret bound for OGD). *Under Assumption 1, 2 and 3, online gradient descent (OGD) with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$ for $t \in [T]$ guarantees:*

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2}GD\sqrt{T}.$$

Stochastic Gradient Descent

Theorem 7 (Convergence of SGD). *Suppose the domain $\mathcal{X} \subseteq \mathbb{R}^d$ has a diameter $D > 0$, and the noisy gradient oracle is unbiased and variance bounded by G^2. SGD with step size $\eta_t = \frac{D}{G\sqrt{t}}$ guarantees*

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] \leq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \frac{3GD}{2\sqrt{T}},$$

where $\bar{\mathbf{x}}_T \triangleq \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t$ is the output of the SGD algorithm.

- We define the linear function $h_t(\mathbf{x}) \triangleq \mathbf{g}_t^\top \mathbf{x} = \mathbf{g}(\mathbf{x}_t)^\top \mathbf{x}$ and run OGD on $\{h_t\}_{t=1}^T$.

- Note that function h_t ***depends*** on the decision $\mathbf{x}_t$, which actually reveals that OGD regret can hold even against ***adaptive adversary***.

More bits of OGD

- We define the linear function $h_t(\mathbf{x}) \triangleq \mathbf{g}_t^\top \mathbf{x} = \mathbf{g}(\mathbf{x}_t)^\top \mathbf{x}$ and run OGD on $\{h_t\}_{t=1}^T$.

- Note that function h_t **depends** on the decision $\mathbf{x}_t$, which actually reveals that OGD regret can hold even against **adaptive adversary**.

At each round $t = 1, 2, \cdots$

   (1) the player first picks a model $\mathbf{x}_t \in \mathcal{X}$;

   (2) and simultaneously environments pick an online function $f_t : \mathcal{X} \to \mathbb{R}$;

   (3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

oblivious adversary

examination

adaptive adversary

interview

The "simultaneously" requirement can be sometimes not necessary!

OGD for full-info OCO can handle the case when online functions depend on $\mathbf{x}_t$!

Online-to-Batch Conversion

- An alternative way to solve statistic learning:
  - use the data in a sequential way
  - run any online algorithm minimizing the regret
  - return the final model as the average

Algorithm 1 Online-to-Batch Conversion

Input: Data $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_T, y_T)\}$ i.i.d. sampled from the distribution $\mathcal{D}$, a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, an online learning algorithm $\mathcal{A}$

1: **for** $t = 1, \cdots, T$ **do**
2:     let $\mathbf{w}_t$ be the output of algorithm $\mathcal{A}$ for this round
3:     Feed algorithm $\mathcal{A}$ with loss function $f_t(\mathbf{w}) = \ell(\mathbf{w}^\top \mathbf{x}_t), y_t)$
4: **end for**
5: **return** $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t$

Online-to-Batch Conversion

Theorem 2 (Online-to-Batch Conversion). *If the risk $R(\mathbf{w})$ is convex w.r.t. $\mathbf{w}$ with a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, and the data $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_T, y_T)\}$ are i.i.d. sampled from the distribution $\mathcal{D}$, then with probability at least $1-\delta$, the excess risk of the output of Algorithm 1 satisfies*

$$R(\bar{\mathbf{w}}) - \min_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w}) \leq \frac{\text{Regret}_T}{T} + 2\sqrt{\frac{2\ln(2/\delta)}{T}}$$

where $R(\mathbf{w}) \triangleq \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}}[\ell(h(\mathbf{w}; \mathbf{x}), y)]$ is the expected risk, and $\text{Regret}_T \triangleq \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w})$ is the regret of the online algorithm $\mathcal{A}$ after T rounds.

Concentration Inequalities

Lemma 1 (Hoeffding's inequality). *Let* $X_1, \ldots, X_T \in [-B, B]$ *for some* $B > 0$ *be independent random variables such that* $\mathbb{E}[X_t] = 0$ *for all* $t \in [T]$, *then for all* $\delta \in (0, 1)$,

$$\Pr\left[\sum_{t=1}^{T} X_t \geq B\sqrt{2T \ln \frac{1}{\delta}}\right] \leq \delta$$

Lemma 2 (Azuma's inequality). *Let* $X_1, \ldots, X_T \in [-B, B]$ *for some* $B > 0$ *be a martingale difference sequence (i.e.,* $\forall t \in [T], \mathbb{E}[X_t \mid X_{t-1}, \ldots, X_1] = 0$ *), then* $\forall \delta > 0$,

$$\Pr\left[\sum_{t=1}^{T} X_t \geq B\sqrt{2T \ln \frac{1}{\delta}}\right] \leq \delta$$

Online-to-Batch Conversion

Theorem 2 (Online-to-Batch Conversion). *If the risk $R(\mathbf{w})$ is convex w.r.t. $\mathbf{w}$ with a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, and the data $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_T, y_T)\}$ are i.i.d. sampled from the distribution $\mathcal{D}$, then with probability at least $1-\delta$, the excess risk of the output of Algorithm 1 satisfies*

$$R(\bar{\mathbf{w}}) - \min_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w}) \leq \frac{\text{Regret}_T}{T} + 2\sqrt{\frac{2\ln(2/\delta)}{T}}$$

where $R(\mathbf{w}) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h(\mathbf{w};\mathbf{x}), y)]$ is the expected risk, and $\text{Regret}_T \triangleq \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w})$ is the regret of the online algorithm $\mathcal{A}$ after T rounds.

Proof Sketch.
$$R(\hat{\mathbf{w}}) \overset{\text{Jensen's inequality}}{\leq} \frac{1}{T}\sum_{t=1}^{T} R(\mathbf{w}_t) \overset{\text{Azuma's inequality}}{\leq} \boxed{\frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{w}_t)} + \sqrt{\frac{2\ln(2/\delta)}{T}}$$

$$\updownarrow \frac{\text{Regret}_T}{T}$$

$$R(\mathbf{w}^*) + \sqrt{\frac{2\ln(2/\delta)}{T}} \overset{\text{Hoeffding's inequality}}{\geq} \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{w}^\star) \geq \boxed{\frac{1}{T}\min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w})}$$

Online-to-Batch Conversion

Proof.

$$R(\hat{\mathbf{w}}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h(\hat{\mathbf{w}};\mathbf{x}),y)\right]$$

$$= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell\Big(h(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t;\mathbf{x}),y\Big)\right]$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h(\mathbf{w}_t;\mathbf{x}),y)\right] \qquad \text{(Jensen's inequality)}$$

$$= \frac{1}{T}\sum_{t=1}^{T}f_t(\mathbf{w}_t) + \sqrt{\frac{2\ln(2/\delta)}{T}} \qquad \begin{array}{l}\text{(Azuma's inequality}\\ \text{with } X_t = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h(\mathbf{w}_t;\mathbf{x}),y)\right] - f_t(\mathbf{w}_t))\end{array}$$
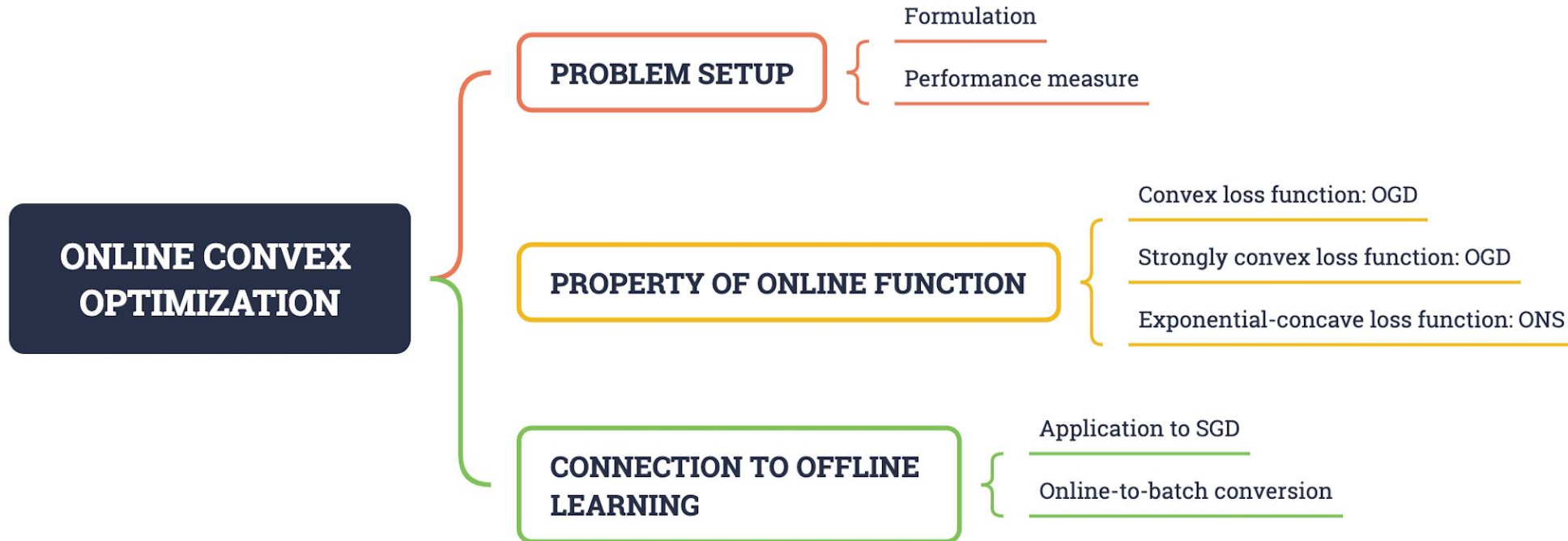
Online-to-Batch Conversion

Proof.

$$R(\hat{\mathbf{w}}) \leq \frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{w}_t) + \sqrt{\frac{2\ln(2/\delta)}{T}}$$

$$= \min_{\mathbf{w} \in \mathcal{W}} \frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{w}) + \frac{\text{Regret}_T}{T} + \sqrt{\frac{2\ln(2/\delta)}{T}} \qquad \text{(definition of regret)}$$

$$\leq \frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{w}^\star) + \frac{\text{Regret}_T}{T} + \sqrt{\frac{2\ln(2/\delta)}{T}}$$

$$\leq R(\mathbf{w}^\star) + \frac{\text{Regret}_T}{T} + 2\sqrt{\frac{2\ln(2/\delta)}{T}} \qquad \begin{array}{l}\text{(Hoeffding's inequality} \\ \text{with } X_t = f_t(\mathbf{w}^\star) - R(\mathbf{w}^\star))\end{array}$$

$\square$

Summary

Q & A

Thanks!